MONITORING AND CROSS-CHECKING AUTOMATION: DO FOUR EYES SEE MORE THAN TWO?

Cymek, D.H.*, Jahn, S.* & Manzey, D.H.*, Technische Universität Berlin
*All authors contributed equally to the manuscript and are displayed in alphabetic order.

The present study addresses effects of human redundancy on automation monitoring and cross-checking. Thirty-six participants performed a multi-task, consisting of three subtasks that mimic basic work demands of operators in a control room of a chemical plant. One of the tasks was to monitor and cross-check a highly reliable and safety-critical automated process. Participants were randomly assigned to two groups: (1) "Non-redundant": participants worked on all tasks alone as the only responsible operator. (2) "Redundant": participants were informed that a second crewmate would work in parallel on the automation monitoring task and that they both were responsible for ensuring safe operation of the automation. Results provide evidence for social loafing effects in automation cross-checking. Participants working redundantly with another crewmate were found to cross-check the automation significantly less than participants, who were working alone. Even if the combined team performance of the participants working in the redundant condition was considered, the number of cross-checks did not significantly differ from the performance in the non-redundant condition. This result suggests that human redundancy can induce social loafing effects which fully compensate a possible reliability gain intended to be achieved by this measure. It challenges the often stated assumption that "four eyes see more than two" and shows that human redundancy does not necessarily lead to enhanced safety in automation monitoring.

INTRODUCTION

In our industrialized society, fast data transmission, fast supply of energy, and fast transportation of goods and humans have become inevitable. Only with the help of highly specialized and complex technology and the cooperation of humans and machines, modern society is able to grow expeditiously. Technological inventions are pushing machines in various work fields to automate and amend human task fulfillment. Therefore in many fields humans are nowadays assisted by more or less complex technology.

The remaining human responsibilities in interaction with highly automated systems have been characterized as *supervisory control* (Sheridan & Parasuraman, 2005). In this concept the human tasks include constant monitoring and cross-checking of fully automated systems, which work highly reliable (though often not perfect). In case the system reaches its capability limits, it is the human task to intervene by e.g. switching to manual operation.

With the demand of monitoring and cross-checking automated systems, human attention and vigilance is challenged. As has often been stated, humans principally are not very well suited for monitoring tasks and may suffer from decrements of vigilance over time (Mackworth, 1948). Another issue often mentioned in this context is the issue of *complacency*. Complacency has been defined as a sort of substandard automation monitoring which, as a performance consequence can lead to poorer detection of system malfunctions under automation control compared with manual control (Parasuraman & Manzey, 2010). It represents a risk factor specifically when humans have to supervise a highly reliable system (Parasuraman, Molloy & Singh, 1993). Parasuraman and Manzey (2010) believe that complacency manifests itself in a malevolved attention allocation strategy (e.g. reduced cross-checking behavior) which results from a learning process in interaction with highly reliable systems referred to as *learned carelessness* (Luedtke & Moebus, 2005). Both, a vigilance decrement as well as complacency effects can cause a loss of *situation awareness* in interaction with an automated system. In case of automation failures, a reduced situation awareness, in turn, can lead to *errors of commission* (following automation directive although it is false) or *errors of omission* (failure to respond to system irregularities or events when automated devices fail to detect or indicate them) (Skitka, Mosier, Burdick & Rosenblatt, 2000). Because of its safety-relevance in all domains where human operators have the role of supervising controllers of automated processes, countermeasures have to be taken to mitigate the risks arising from vigilance decrements and/or complacency effects.

One specific countermeasure, particularly recommended for use in safety-critical organizations (e.g. nuclear power plant, chemical plant, airplane), is raising the number of persons, who monitor and cross-check the automated system. This principal is called *redundancy* and has been used for a long time as an engineering tool to enhance the overall reliability and safety of technical systems. Redundancy theory demonstrates that the implementation of redundant technical components, if independent and connected in parallel manner, can lead to rapid increases in overall system reliability (Sagan, 2004). For example, if the probability of a failure for one single component is 1/10, it is already reduced to $(1/10)^2$ if two independent components are redundantly used and $(1/10)^3$ when a third redundant component is added. Many security analysts have, therefore, advised the widespread deployment of redundancy as a key requirement of *high reliability organizations* (*HRO*). However, in case of the so called *human redundancy* the relation between redundant components, i.e. redundantly working operators and safety enhancement is not

as straightforward as it is in purely technical redundant systems (Clark, 2005; Conte & Jacobs, 1997). The awareness of each other and the knowledge of the task being accomplished multiple times violate the precondition of component independence. As a consequence, individuals may reduce their individual effort - an effect that is well documented and termed *social loafing* in social psychology (Latané, Williams & Harkins, 1997). Karau and Williams (1993) believe that individuals start loafing because they are only willing to exert effort on a collective task to the degree that they expect their efforts to be instrumental in obtaining valued outcomes. One-hundred percent reliable automation cross-checking is possible in most cases with only one motivated and diligent person. Having this in the back of one's mind, individuals may reduce automation monitoring if this task is performed collectively compared to situations where only one person is in charge of the task. Sagan (2004) even supposed that introducing human redundancy in complex technological facilities might actually raise safety issues instead of lowering them. He displayed that when the reliability of each new person in a redundant system leads to a reduction of the individual reliability due to social loafing effects by about 15%, the overall team reliability theoretically will improve for redundant crews of two or three operators but starts to decline and to become less than the reliability of a non-redundant single-operator system if the crew of operators becomes larger.

Thus, a thorough investigation of possible social loafing effects in a redundantly performed cross-checking task is necessary before recommending human redundancy as a design element that mitigates reduced cross-checking due to complacency and/or vigilance decrement.

The present study addresses this issue and extends the research from Skitka et al. (2000), Domeinski, Wagner, Schoebel and Manzey (2007) as well as Manzey, Boehme and Schoebel (2013). Skitka et al. (2000) investigated whether the tendency towards errors of omission and commission were ameliorated when two instead of one single decision maker is monitoring system events. No difference between team performance and solo performance was found with respect to committing commission or omission errors, suggesting that no safety gain might be achieved by human redundancy. However, in this study no direct measures of individual monitoring behavior were taken which made any interpretation in terms of possible social loafing effects in the team condition difficult. Domeinski et al. (2007) and Manzey et al. (2013) compared the actual cross-checking behavior of participants who were believed to work alone or redundantly with another crew mate on an automation monitoring task which was part of a complex multi-task environment. Individuals working in the redundant condition were found to reduce the number of cross-checks of the automation significantly, compared to individuals working alone. This also led to a higher risk of committing errors of omission for participants working in the redundant condition. Although these results provide clear evidence for social loafing effects on the individual level, they do not necessarily challenge the concept and benefits of human redundancy, as none of these studies directly compared the combined team performance of redundant working dyads with the performance of single operators. Obvious ceiling effects in the condition where participants worked as single operators made any such evaluation difficult.

Based on these considerations and using a multi-task environment closely resembling the one used in previous research (e.g. Manzey et al., 2013), the present experiment addresses three different aspect. First, we wanted to replicate the earlier findings of social loafing effects in automation monitoring introduced by human redundancy. Second, we wanted to investigate to what extent social loafing effects are influenced by time-on-task. Third, and most important, we wanted to investigate whether the combined monitoring performance of a redundant dyad would still be better than the performance of a single person. This latter effect would provide strong arguments for considering human redundancy an effective safety countermeasure for known issues in human-automation interaction even in case of social loafing effects.

METHOD

Participants

The sample used for this study consisted of 36 participants of whom 35 were students (23 female), aged 20-29 years (mean age: 25.17 years). The participants had no prior experience with the task. They were compensated for their participation with 15 Euro or three student credit hours.

Apparatus and Task

For the present experiment a PC-based laboratory multi-task environment was used (Multi-Task Operator Performance Simulation for Redundancy Research, M-TOPS-R). It features three subtasks which are designed to mimic basic work demands of operators in a control room of a chemical plant. The user-interface is shown in Figure 1.

Resource ordering task (ROT). This task is shown in the upper left and basically represents a mental arithmetic task. Participants are instructed to assure the availability of required chemicals in order to keep the chemical process running. For this purpose the actual and the set values of different catalysts are presented, each for only ten seconds. Participants have to calculate the differences and type them into the ordering field. To submit they have to click on the order icon. After an order has been sent or after the ten seconds have passed a new task is presented after an interval of two to five seconds.

Coolant exchange task (CET). This task is displayed in the upper right. Participants have to exchange the fluids in two vessels of a cooling system. Thereto different valves have to be opened and closed in a defined sequence to drain used fluid (green fluid) and to refill the vessels with fresh fluid (blue fluid). The speed with which the vessels are filled and emptied varies to avoid highly routinized workflows. The minimal time for a complete exchange-cycle therefore ranges from 14 to 38 seconds.

Monitoring task (MOT). This task is displayed in the lower right. Participants here have supervisory control over a system that autonomously analyses and controls the processes in different reaction chambers. Participants receive information about the assumed current state of the reaction chamber (e.g.

"temperature high"; "process running") and the action the automation wants to initiate accordingly (e.g. "reduction of temperature"; "no action needed"). Participants are made believe that the automation would work quite reliable but not perfect. Therefore automation performance needs to be cross-checked. To do so participants are asked to verify the automation by double-checking the raw data of the simulated process (e.g. temperature, heat distribution, pressure, valve setting). To view each parameter's raw data, participants have to left-click on the corresponding button. Each possible state in a given chamber is defined by two conditions. Therefore full automation verification always comprises the inspection of two different parameters. In case an automation failure is identified (e.g. indication of "temperature high" when temperature is in normal range), participants have to initiate manual control (left-click on "manual" icon) for the given chamber. If no cross-checking is initiated within five seconds, the current reaction chamber disappears and the next chamber occurs after two to five seconds. If a participant decides to cross-check the automation a time credit of three seconds is warranted for every parameter clicked upon.

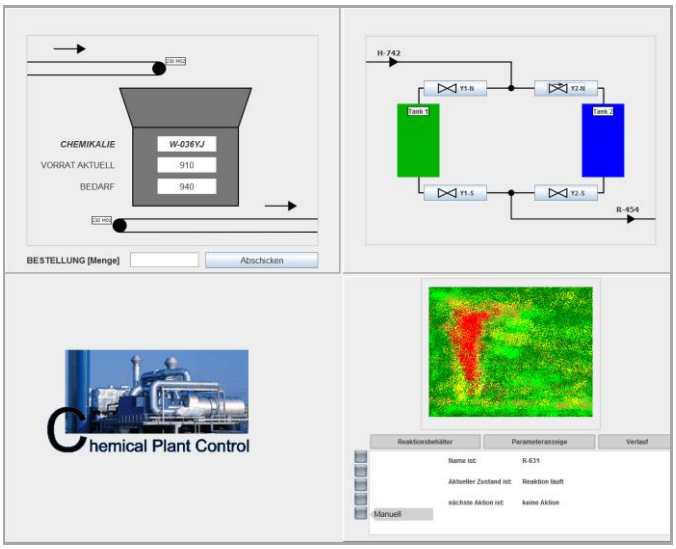

Figure 1. User interface of Multi-Task Operator Performance Simulation for Redundancy Research (M-TOPS-R). Upper left: resource ordering task (ROT); upper right: coolant exchange task (CET); lower right: monitoring task (MOT).

Design

For the present study a 2 (*working condition*) x 12 (*block*) mixed factorial design was used. The first factor *working condition* represented a between-subjects factor with the two levels defined as (1) non-redundant work and (2) redundant work. In the *non-redundant* working condition it was stated that the participants would work on all three tasks alone. In the *redundant* condition participants were instructed that they would be the only responsible operator for the coolant exchange (CET) and resource ordering task (ROT), but would handle the monitoring task (MOT) as part of a two-person crew, i.e. would perform this subtask redundantly with the crewmate sitting next to them. The instructions further explained that only their team performance would be evaluated, i.e., if one of them detected an automation error their partner wouldn't be alerted to it but the trial would count as correctly checked for both of them. Participants were randomly assigned to one of these two conditions.

The second experimental factor *block* represented a within-subjects factor and was introduced to consider possible time-on-task effects. This factor depicts the changes in performance over time for all tasks via 12 successive blocks, each consisting of 12 monitoring trials (four minutes on average). The automation worked without a single failure in order to increase the participants' trust and to replicate the very high reliability of most automated systems which can be assumed to induce processes of learned carelessness over time.

Dependent variables

All performance measures were derived from a participant-specific log-file that recorded all actions done during the experiment.

The main focus of the experiment was to evaluate the effects of the experimental conditions on MOT performance on individual as well as on team level. Thereto two separate measures were considered, one to assess the individual monitoring performance in both, the non-redundant and redundant conditions, and one to assess the team monitoring performance in the redundant condition.

Individual monitoring performance. In order to assess how carefully the individual participants verified the automated diagnoses by cross-checking the available raw data, the number of trials where the automation was fully cross-checked was computed for each individual in each condition and each block. Only events where participants accessed both parameters necessary to cross-check a given diagnosis and action indicated by the automation for a given chamber were counted as "fully cross-checked". A maximum of twelve cross-checks was possible per block.

Team monitoring performance. To assess the automation monitoring and cross-checking performance on team level in the redundant condition, the number of full cross-checks performed by the two-person teams was counted. For this purpose, a trial was counted as a "full cross-check" trial when at least one of the two team members in the redundant working condition had cross-checked both parameters necessary to verify the diagnosis and the action indicated by the automation. As for individual participants, also teams could achieve a maximum of twelve cross-checks per block.

Furthermore, subjective measures were collected to assure the effect of the experimental manipulation on participants' perception and behavior. Therefore participants' feelings of responsibility, liability and motivation, as well as their assessment of their own and their partner's performance were assessed using questionnaires. Participants indicated their agreement to the following statements on a 4-point Likert scale ranging from 1 = *a little* to 4 = *very much*: "I felt reliable/liable/was motivated to do well at the [insert task]" and "I/ my partner performed well at [insert task]/ compared to me". Lastly, the physical condition before and after the experiment was assessed with the Stanford Sleepiness Scale (Hoddes, Dement & Zarcone, 1972) as well as with one self-developed

item, "My current performance capability is at _____% of my maximum performance capability," to ensure that no confounding effects existed.

Finally, also performance measures for the two other subtasks were collected but will not be reported here.

Procedure

Four working stations had been set up in pairs next to one another, divided by moveable walls to prevent gazes to the other workstations. Participants were further instructed to not talk to each other.

At the beginning, participants filled out questionnaires about their demographics and current performance state (i.e. physical condition). Then, a general instruction to M-TOPS-R was provided, followed by a detailed written explanation of the three subtasks including the automation verification procedure to be used in the MOT. Each task was trained separately first. A practice trial followed in which participants had to do all three tasks simultaneously for six minutes.

The manipulation of the two working conditions was done with differing instructions. In the non-redundant condition participants were told that they would work on independent work stations. In the redundant conditions partners had been assigned from the instructor and were introduced to each other as being team partners at the beginning of the experiment. To further support the illusion that the participants in the redundant condition were working in teams they had to wait for their partners to enter the IP-address of their work station into the system to be able to start the practice trial. After a short break, the practice trial was succeeded by the actual data collection of the experimental task, which lasted about 48 minutes (144 trials divided into 12 blocks á four minutes). No breaks were provided between the blocks, neither was the block structure made transparent for the participants.

After the experiment participants were asked to fill out the final questionnaire and to state their current physical condition. The session concluded with a debriefing about the objectives of the experiment and the faked redundant condition before participants received their compensation.

RESULTS

Manipulation Check and Subjective Data

To assure that the experimental manipulation regarding the two working conditions, non-redundant and redundant, really worked, participants were asked to vote their agreement to the statement "I was solely responsible for the [insert task]" on a 4-point Likert scale. For the MOT the agreement was supposed to be high in the non-redundant group and low in the redundant group. The results show that the manipulation can be considered successful, with participants rating their agreement higher (*Mdn* = 4) for the non-redundant condition than for the redundant condition (*Mdn* = 1.5). The difference between both groups was significant with $U_{MOT} = 63$, $p < .01$.

Furthermore, no significant difference between the two experimental groups emerged with regard to motivation, feeling of liability and subjective assessment of performance.

Lastly, no obvious relationship between physical condition and performance was found.

Performance Measures

Individual monitoring performance. The number of completely performed cross-checks per block was considerably smaller in the redundant condition ($M = 84.56$) as in the non-redundant one ($M = 112.28$). A 2 (*working condition*) x 12 (*block*) ANOVA revealed a significant difference, $F(1, 34) = 4.89$, $p = .03$, $\eta_p^2 = .13$. Because of violations of the sphericity assumption, degrees of freedom for the effect of *block* were corrected according to Greenhouse-Geisser. The analysis revealed a main effect for the factor *block*, $F(7.15, 243.13) = 2.42$, $p = .02$, $\eta_p^2 = .07$, indicating that monitoring performance differed across the 12 blocks. In addition, a significant interaction effect *working condition* x *block* emerged, $F(11, 374) = 2.79$, $p = .002$, $\eta_p^2 = .08$. As becomes evident from Figure 2, the mean number of full cross-checks slightly decreased over successive blocks for participants working with a partner (redundant: white circles), while the number of cross-checks slightly increased for the participants working alone (non-redundant: black circles).

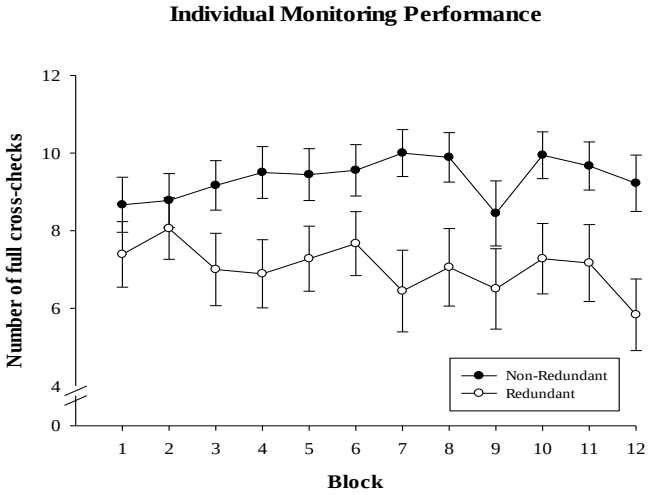

Figure 2. Mean numbers and standard errors of full cross-checks in the monitoring task over 12 blocks. The non-redunant graph (black circles) shows the performance of the participants working alone, the redundant graph (white circles) depicts the performance of individual persons during the work with a team partner.

Team monitoring performance. The data for the team performance revealed that, summed across blocks, more full cross-checks in the monitoring task were performed by both team members taken together ($M = 125.00$) than by the participants working alone ($M = 112.28$). This difference (team vs. non-redundant), however, failed to reach significance, $F(1, 34) = 1.91$, $p = .18$, $\eta_p^2 = .05$. Because of violations of the sphericity assumption, degrees of freedom for the effect of the factor *block* were corrected according to Greenhouse-Geisser. The analysis revealed a main effect for this factor, $F(6.75, 229.63) = 3.52$, $p < .01$, $\eta_p^2 = .1$. In addition, a significant interaction effect *condition* (team vs. non-redundant) x *block* emerged, $F(11, 374) = 3.07$, $p < .01$, $\eta_p^2 = .08$. Figure 3 shows that the redundant teams (white circles) outperformed

the non-redundant working participants (black circles) during the first two blocks. However, due to an increase of individual performance in the non-redundant condition, this benefit was largely reduced and almost vanished for the remaining blocks but block #9.

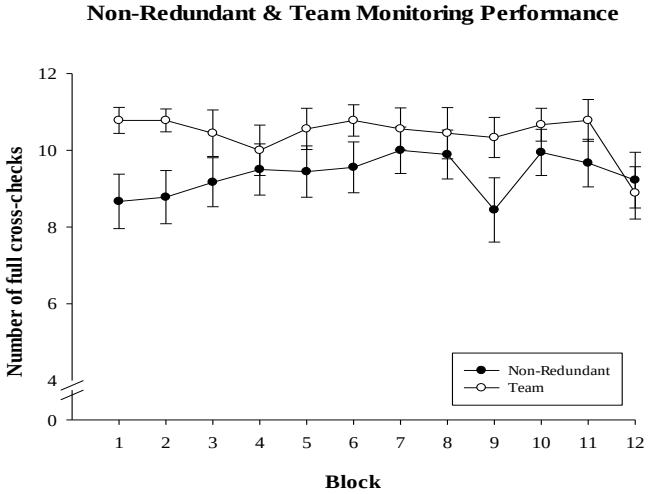

Figure 3. Mean numbers and standard errors of full cross-checks in the monitoring task over 12 blocks. The non-redundant graph (black circles) shows the performance of participants working alone and the team graph (white circles) the combined performance of two persons working in a team.

DISCUSSION

The results of the present experiment again create doubts on whether human redundancy is indeed a design element that can mitigate known risks of human automation monitoring. This can be concluded from three different findings:

First, the result of the present experiment confirms earlier results reported by Domeinski et al. (2007) and Manzey et al. (2013), which indicate that individuals working redundantly with a team partner on an automation monitoring task reduce their effort to cross-check the automation compared to participants working alone. This finding once again provides evidence for the risk of social loafing effects in redundant teams.

Second, while the number of automation cross-checks slightly increased over blocks (i.e. time-on-task) for participants working in the non-redundant condition, possibly indicating effects of learning and routinization, a slight decrease of individual performance was found in the redundant condition. This opposing trend further questions the safety gain of human redundancy in particular when redundant work conditions persist over longer periods. In the present study participants worked on the multi-task only for 48 minutes. Therefore it should be further investigated whether the declining trend pursues with longer working times.

Third, although a small advantage of combined team performance as compared to individual performance was observed on a descriptive level, the number of full cross-checks performed by teams was not significantly different from the performance of participants in the non-redundant condition. Thus the overall system reliability was not enhanced through redundant task completion. This confirms earlier results from studies of Skitka et al. (2000) and Mosier et al. (2001) who also did not find obvious performance benefits of teams compared to individuals when considering the risk of commission and omission errors during the interaction with an automated system. Sagan (2004) predicted that even unreliable components, if independent and parallel, lead to a rapid increase of overall system reliability when combined to redundant systems. Since humans working in teams are aware of each other the independence assumption is violated and the rapid increase in safety is uncertain. The present findings emphasize that social loafing effects in teams can largely compensate reliability gains of human redundancy and, hence, render reliability of teams not considerably better than the reliability of a single person. As a consequence, the advice of safety experts to deploy redundancy as a key requirement of high reliability organizations (HRO) may be appropriate for technical components, but questionable for the case of human redundancy.

REFERENCES

Clarke, D. M. (2005). Human redundancy in complex, hazardous systems: A theoretical framework. *Safety Science*, *43*(9), 655–677.

Conte, J. M., & Jacobs, R. R. (1997). Redundant Systems Influences on Performance. *Human Performance*, *10*(4), 361–380.

Domeinski, J., Wagner, R., Schoebel, M., & Manzey, D. (2007). Human Redundancy in Automation Monitoring: Effects of Social Loafing and Social Compensation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(10), 587-591.

Hoddes, E., Dement, W., & Zarcone, V. (1972). The development and use of the Stanford Sleepiness Scale (SSS). P*sychophysiology*, *9*(1), 150.

Karau. S. J., & Williams, K. D. (1993). Social Loafing: A meta-analytical review and theoretical integration. *Journal of Personality and Social Psychology*, 65, 681-706.

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, *37*(6), 822-832.

Luedtke, A., & Moebus, C. (2005). A case study for using a cognitive model of learned carelessness in cognitive engineering. In G. Salvendy (Ed.), *Proceedings of the 11th International Conference of Human-Computer Interaction*. Mahwah, NJ: Erlbaum. Retrieved from http://www.lks.uni-oldenburg .de/download/abteilung/Luedtke_Moebus_crv.pdf

Mackworth, N.H. (1948). The breakdown of vigilance during prolonged visual search, Quarterly *Journal of Experimental Psychology*, 1, 6-21.

Manzey, D., Boehme, K., & Schoebel, M. (2013). Human Redundancy as Safety Measure in Automation Monitoring. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 369-373.

Mosier, K. L., Skitka, L. J., Dunbar, M., & McDonnell, L. (2001). Aircrews and Automation Bias: The Advantages of Teamwork?. *The International Journal of Aviation Psychology*, *11*(1), 1-14.

Parasuraman, R., Molloey, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology*, *3*(1), 1-23.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(3), 381-410.

Sagan, S. D. (2004). The problem of redundancy problem: why more nuclear security forces may produce less nuclear security. *Risk analysis*, *24*(4), 935-946.

Sheridan, T. B., & Parasuraman, R. (2006). Human-automation interaction. In R. S. Nickerson (Ed.), *Reviews of Human Factors and Ergonomics*, *1*(1), 89-129.

Skitka, L. J., Mosier, K. L., Burdick, M., & Rosenblatt, B. (2000). Automation bias and errors: are crews better than individuals?. *The International Journal of Aviation Psychology*, *10*(1), 85-97.